Published online 11 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D411–D419***
doi:10.1093/nar/gkq1105

A series of PDB related databases for everyday needs

Robbie P. Joosten[1], Tim A.H. te Beek[2], Elmar Krieger[2], Maarten L. Hekkelman[2], Rob W.W. Hooft[3], Reinhard Schneider[4], Chris Sander[5] and Gert Vriend[2,*]

[1]Department of Biochemistry, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, [2]Centre for Molecular and Biomolecular Informatics, NCMLS 260, Radboud University Nijmegen Medical Centre, Geert Grooteplein 26, 6525 GA Nijmegen, [3]Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands [4]European Molecular Biology Laboratory – Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany and [5]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA

Received September 14, 2010; Revised October 15, 2010; Accepted October 18, 2010

ABSTRACT

The Protein Data Bank (PDB) is the world-wide repository of macromolecular structure information. We present a series of databases that run parallel to the PDB. Each database holds one entry, if possible, for each PDB entry. DSSP holds the secondary structure of the proteins. PDBREPORT holds reports on the structure quality and lists errors. HSSP holds a multiple sequence alignment for all proteins. The PDBFINDER holds easy to parse summaries of the PDB file content, augmented with essentials from the other systems. PDB_REDO holds re-refined, and often improved, copies of all structures solved by X-ray. WHY_NOT summarizes why certain files could not be produced. All these systems are updated weekly. The data sets can be used for the analysis of properties of protein structures in areas ranging from structural genomics, to cancer biology and protein design.

INTRODUCTION

The Protein Data Bank (PDB) (1–4) is the world-wide repository of macromolecular structures solved by X-ray diffraction, NMR, or (cryo-)electron microscopy. More than 67 000 entries in the PDB (summer 2010) are a true treasure trove for scientists in fields such as protein engineering, human genetics, drug design, molecular biology, biochemistry, etcetera. In protein engineering, for example, one often needs to know that whether a residue is conserved, and if not, which residue types are observed at the equivalent positions in related proteins. In human genetics, one often wonders where an observed disease causing mutation is located in the structure relative to the active site, the DNA binding site, a multimer interface or another functionally important site. Such questions, and many more, normally can be answered if the structure of the protein at hand is available, and if a lot of additional data and tools are available. In protein structure bioinformatics it is common practice to use molecular visualization software, the UniProtKB/Swiss-Prot (5) file, a multiple sequence alignment, a report about the quality of the structure used, articles and many other types of information.

We maintain a series of databases, Web servers and Web services to aid the scientists with their macromolecular structure-based research. About 75 Web servers that take PDB files as input are available at http://swift.cmbi.ru.nl/ (unpublished results), and we recently described the first series of 50 Web services (6) that act on PDB files. These facilities are used hundreds to thousands of times per day, but in some cases it makes better sense to pre-store the results rather than to generate them on-the-fly. One reason is that a result may be used frequently. In that case, generating the same result over and over is a waste of CPU time on the side of the server and waiting time on the side of the user. Another reason is that some PDB derived results take too much time to generate to be used as a service. Creating a typical entry for the PDB_REDO database (7), for instance, takes several hours. The most important reason to store results in databases instead of generating them when they are needed is that it allows for quick data mining. Say, we want a list of all PDB entries that contain threonine residues with inversed chirality at their Cβ atoms. Checking all threonines in the PDB will take hours or even days, checking all the PDBREPORT (8) records

*To whom correspondence should be addressed. Tel: +31 24 36 19390; Fax: +31 24 36 19395; Email: vriend@cmbi.ru.nl

for this specific problem will only take minutes, and with a pre-indexed version of PDBREPORT this list can be retrieved in seconds.

We describe several databases that can be used to obtain insight in the many aspects of a specific protein, but can also help to select data sets for (structural) analysis, to find properties of proteins in general, or find suited test sets to create, test and optimize new methods in structural biology research.

At http://swift.cmbi.ru.nl/gv/facilities/ an overview of all systems mentioned (and a few more) is given, and pointers are provided to extensive documentation that includes help for downloading whole databases.

DATABASES

A short summary of our databases, their purpose and their locations is given in Table 1. The first four databases listed in Table 1 provide PDB file annotation in terms of structure, sequence and quality (and the improvability thereof). The next three are aimed at data set selection and are partly derived from the first four databases. The final database provides information about the entries of the other databases, or rather about the entries missing from these databases.

The secondary structure of proteins is an important aspect in many fields of bioinformatics. A simple Google search for the exact string 'secondary structure prediction server' gives more than 70 000 hits and new methods to predict protein secondary structure are still published regularly (9–13). This might seem a bit surprising because there are not that many biological questions that require knowledge of a protein's secondary structure to be answered, but in practice the secondary structure of a protein is an important tool for classification and comparison purposes (see for examples the CATH (14) and SCOP (15) protein classification databases).

Table 1. List of available databases with a short summary of contents and location

Database name	Database description
DSSP	Secondary structure of proteins http://swift.cmbi.ru.nl/gv/dssp/
HSSP	Multiple sequence alignments of UniProtKB against PDB files http://swift.cmbi.ru.nl/gv/hssp/
PDBREPORT	Reports about errors and anomalies in macromolecules http://swift.cmbi.ru.nl/gv/pdbreport/
PDB_REDO	Re-refined PDB files solved by X-ray crystallography http://www.cmbi.ru.nl/pdb_redo/
PDBFINDER	Searchable summaries of PDB file information ftp://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder/
PDBFINDER2	As PDBFINDER, but with much extra information added ftp://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder2/
PDB_SELECT	Quality-sorted culled lists of protein chains in the PDB http://swift.cmbi.ru.nl/gv/select/
WHY_NOT	Explanation why entries in other databases cannot exist http://www.cmbi.ru.nl/WHY_NOT/

http://swift.cmbi.ru.nl/gv/facilities/ holds both an overview of all systems and detailed information for each of them.

The DSSP software (16) describes the secondary structure of a protein based on its three dimensional structure. Over the years, several alternatives for DSSP have been produced. Looking at DSSP's thousands of citations, and at the fact that today, nearly 30 years after DSSP was written, this software is still distributed on average at least once per week and cited 4–5 times per week, it is safe to state that DSSP is the *de facto* standard in the field of secondary structure determination, and thus also in the field of secondary structure prediction. The DSSP database contains DSSP descriptions for every PDB entry. Figure 1 shows a very small part of a DSSP file with a short explanation.

The concept of residue conservation is highly conserved in many protein structure related research fields and has been mentioned many tens of thousands of times in the literature. A literature search reveals that sequence conservation is used to improve alignments, to score docking solutions, to find functional regions, to cluster residues involved in similar aspects of the protein's function, in drug design, in optimizing HIV drug administration regimes, in evolutionary studies, in the prediction of protein interaction surfaces, in structure-function relation predictions, in secondary structure prediction, in the analysis of crystal contacts and in protein engineering, to mention just a few of the applications. The HSSP [Homology-derived Secondary Structure of Proteins; (17–21)] database holds for each PDB entry a multiple sequence alignment against all UniProt entries that can be aligned against the PDB file's sequence with 5% more confidence than required to infer structural similarity (Figure 2). The sequence variance and the sequence entropy at each position in the protein sequence are given. Together with the alignment, this illustrates the structural and functional importance of each residue in the PDB file.

PDB files are the result of experimental work, and thus are prone to experimental errors. These errors range from administrative mistakes such as violation of nomenclature, through small inaccuracies in bond geometry and small mistakes like wrong side chain rotamers, badly modelled flexible loops, or strange solvent models, all the way to gross errors, a few of which have lead to retractions (e.g. (22)). We have designed the WHAT_CHECK (23–31) software to search for these errors, to list them, quantify them, to try to find their origin and to suggest how to fix them when possible. We ran this software on all PDB files and the resulting reports list about 8.5 million errors, 33.6 million warnings and 17.2 million notes. These reports are available from the PDBREPORT database (8). The WHAT_CHECK reports present the users with 10 sections. The first two sections deal with problems that are detrimental to quality of the validation in the sections that follow. These sections deal with space group related topics, topology determinations, missing atoms, etc. The third section provides a description of the molecule that is informative for the quality; this includes the Ramachandran plot, and the secondary structure as described by DSSP. Further sections deal with occupancies, B-factors, terminal groups, nomenclature issues, elementary geometric features, torsion

```
#  RESIDUE AA STRUCTURE BP1 BP2  ACC
  1    2 A T             0   0   77
  2    3 A T  E     -A   34  0A   21
  3    4 A a  E     -A   33  0A    0
  4    5 A b        -     0   0    0
  5    6 A P  S    S+     0   0   52
  6    7 A S  S  > S-     0   0   48
  7    8 A I  H  > S+     0   0  123
  8    9 A V  H  > S+     0   0   98
  9   10 A A  H  > S+     0   0    6
 10   11 A R  H  X S+     0   0   55
```

B = residue in isolated β-bridge
E = extended strand, in β-ladder
G = 3-helix (3/10 helix)
H = α-helix
I = 5-helix (π helix)
T = hydrogen bonded turn
S = bend

Figure 1. Essentials of the DSSP file. Left: a small part of a secondary structure description. From left to right the columns contain the sequential number of the amino acid, its PDB number, the chain identifier, the amino acid sequence (with paired cysteines replaced by pairs of lower case characters), the actual secondary structure assignment (in red), a description of the type of turn encountered, in case of β-sheets the partner β-strand(s) and the sequential strand identifier, the solvent accessibility of the residue in square Ångströms. The lines further contain information (data not shown) about the geometry of the hydrogen bond(s) that were used to assign the secondary structure, local backbone angles and torsion angles and the coordinates of the Cα. Right: the meaning of the most used (red) column from DSSP files: the secondary structure assignments. Most people convert B, S and T simply into loop (which is a blank in DSSP), sometimes the G is converted into a H, and the I (π helix) is so rare that people tend to just forget about it.

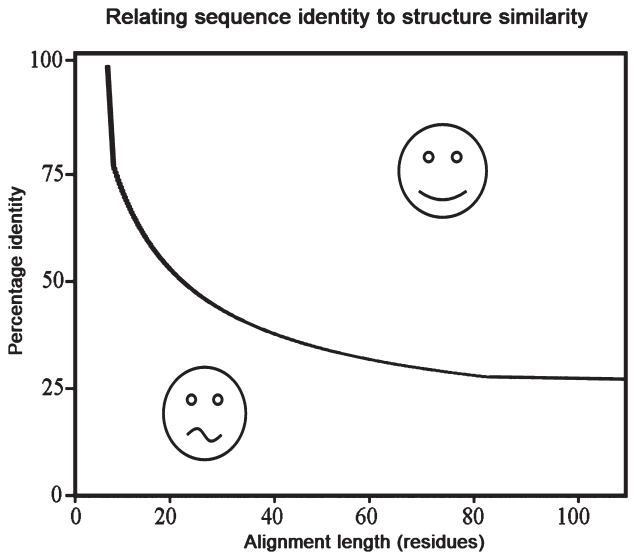

Figure 2. HSSP sequence alignment threshold. The structural homology plot (18) describes at which percentage sequence identity an alignment of a given length is an indication that the aligned proteins have a similar structure. The dark curve gives the cut-off below which no structural similarity can be inferred. This is frequently used in the context of homology modelling: alignments above the curve indicate that it is possible to make a fairly reliable homology model from the aligned template; alignments below the curve mean that a homology model should be handled with care.

angles, proline rings, atomic clashes, threading issues, water molecules and ions and hydrogen bond related topics such as His, Asn, or Gln that need their side chain flipped by 180° to make (better) hydrogen bonds. Each report ends with a summary of the most essential statistics. When used interactively, the WHAT_CHECK software finishes with a set of recommendations for further refinement, but this section is not included in the PDBREPORT database as it is only relevant for the crystallographer solving the structure. Table 2 lists for a few error types their frequency in the PDB.

Table 2. Frequency in the PDB of a few error types listed in the PDBREPORT database

Error	Occurrences in the PDB
Atomic clash	13M
Planarity off by >10σ	141k
Bond length off by >6σ	631k
His, Asn, Gln side chain 'flipped'	486k

The vast majority of structures in the PDB are solved by means of X-ray crystallography. The computational methods to produce a structure model based on the experimental X-ray data have improved dramatically since the beginning of the PDB and still improve today. Additionally, computers can now do in a day what was not even possible in a year in the early 90's, and we understand the biophysical and structural characteristics much better than in the years past. As a result of all this, crystallographers can now build better structure models than ever before. These advances come with a side effect: as new PDB entries improve, older PDB entries, which were solved with older computational methods, start to lag behind in terms of structure quality. To solve this issue, we started applying these new methods to existing, older PDB entries. Using the crystallographic program Refmac (32,33), we re-refined all X-ray structure models in the PDB for which the experimental X-ray data were deposited (34). In this process the fit of the atomic coordinates to the experimental X-ray data is optimized, which improved not only the fit to the experimental data for 67% of the PDB entries, but also the overall quality of structure models as judged by WHAT_CHECK. These updated PDB entries are stored in the PDB_REDO database (7) and can be used for structural biology research exactly as regular PDB files.

The PDB_REDO pipeline is still a topic of intense research in two collaborating groups so that further improvements are expected in the years to come. A recent

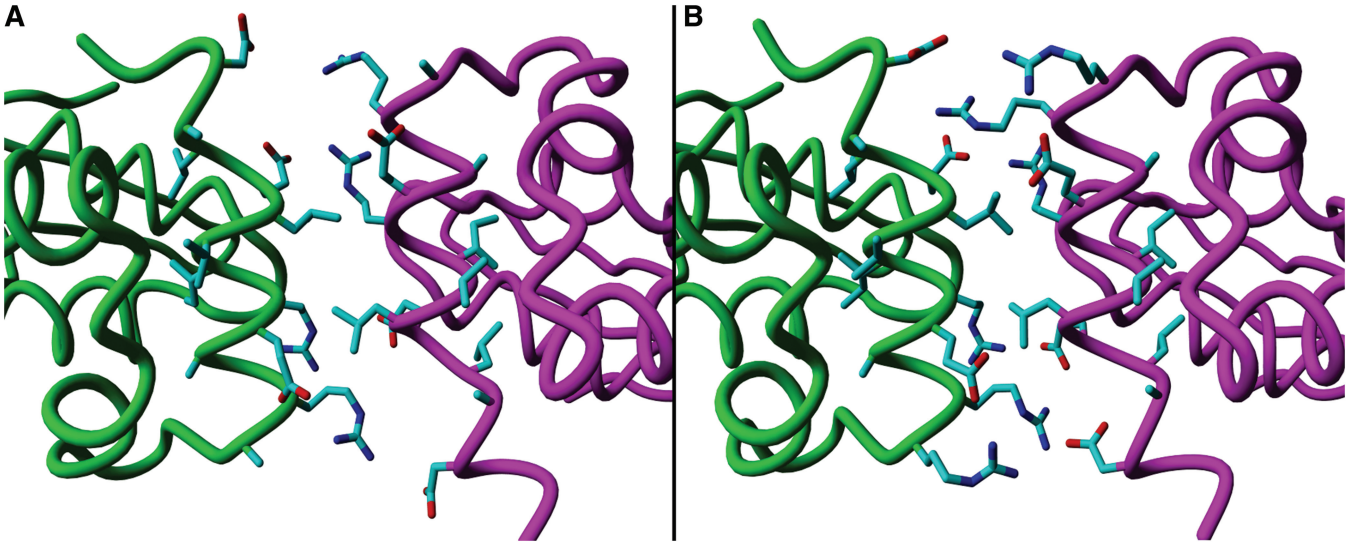

Figure 3. Example of improved structure in PDB_REDO. Detail of the multimer interface of the RuvA-Holliday junction complex. Chain A is shown in magenta and chain F in green. The side chains for helical residues 117–129 are shown as sticks coloured by chemical element. Ionic interactions between the two anti-parallel helices help to form a strong complex. (**A**) The packing interface as deposited in PDB entry 1bvs (35). (**B**) The packing interface from the PDB_REDO optimized structure model. The ionic interactions are similar to the original PDB entry, but because many charged side chains moved inwards, the salt bridges have become shorter. This leads to the conclusion that the interaction between the helices is stronger than previously expected.

improvement is the implementation of new algorithms that optimize the orientation of the peptide planes in the protein backbone, rebuild existing amino acid side chains, build missing ones and optimize hydrogen bonding (unpublished results). This allows PDB_REDO to actively improve structure models instead of relying on the radius of convergence of X-ray refinement software. Over a test set of 4100, PDB entries (deposited from 1996 to 2004) we saw an improvement of the fit of the atomic coordinates with the experimental data for 85% of the test set structures. We are currently updating all PDB_REDO entries to include these new developments. This will be completed by the end of 2010.

PDB entry 1bvs (35) is an example of the extent to which a structure is changed by optimizing it in PDB_REDO. The protein (the RuvA-Holliday junction complex) is an octamer consisting of two tetramers that dimerize by strong ionic interactions between anti-parallel helices (Figure 3a). 1bvs is a relatively low-resolution (3 Å) structure. The methods to refine such structures have improved substantially since the time this structure was solved (in 1998). By employing improved refinement methods, such as TLS refinement (33,36) and local non-crystallographic symmetry restraints, we could improve the fit of the structure model with the data: R-free went down from 31.9 to 28.4%. More importantly, this refinement led to new electron density maps that allowed us to rebuild the side chains at the dimerisation interface (Figure 3b). The rebuilding lead to another improvement of R-free (down to 26.0%) and the rebuilt interface has much better ionic interactions, which is reflected by the residue packing score from WHAT_CHECK (27): the structure moves from a packing score 2σ below the average of a high resolution test set to a score slightly higher than the same set. This means that the case made by the depositor about the nature of the dimerisation interface is now better supported by the updated 3D structure than it was with the original structure.

Considering the substantial size of the aforementioned databases PDB, DSSP, HSSP and PDBREPORT (summing up to 160 GB in the summer of 2010), there is a strong need for a compact summary that can be parsed, searched [e.g. by SRS (37), EBeye (38) or MRS (39)], and analyzed quickly. For this purpose, the PDBFINDER (40) and more recently the PDBFINDER2 databases have been created. Both are actually single flat text files, optimized for minimum size and maximum parsing speed (when compared e.g. to the XML format). While the PDBFINDER (current size 0.16 GB) summarizes the PDB, the PDBFINDER2 includes information from DSSP, HSSP and PDBREPORT, simply added as extra lines (1.1 GB).

As can be seen from the example in Figure 4, the PDBFINDER contains information about the compound (including EC numbers for enzymes), the source, the authors, the experimental method and refinement software (partly manually curated), small molecules (HET-groups), the overall secondary structure content, and a list of all chains (proteins and nucleic acids), including secondary structure content, cysteine bridges and most importantly the sequence. The latter is actually the sequence extracted from the ATOM section of the PDB file, and thus contains only residues for which 3D coordinates are available. This is especially useful for all molecular modelling applications, since other PDB sequence summaries (e.g. the FASTA file generated weekly by the NCBI and commonly used for BLAST searches) are based on the SEQRES section, which includes all residues, even those whose structure could

```
                      PDBFINDER
ID             : 1CRN
Header         : PLANT SEED PROTEIN
 Date          : 1981-07-28
Compound       : crambin
Source         : (crambe hispanica subsp
Source         : organism_taxid: 3721;
Author         : W.A.Hendrickson
Author         : M.M.Teeter
Exp-Method     : X
 Resolution    : 1.50
 R-Factor      : 0.114
Ref-Prog       : PROLSQ
HSSP-N-Align   : 47
T-Frac-Helix   : 0.48
T-Frac-Beta    : 0.09
T-Nres-Prot    : 46
Chain          : A
 Ch-Compnd     : crambin
 Sec-Struc     : 46
  Helix        : 22
   i,i+3       : 3
  Beta         : 4
   Anti-Hb     : 4
 Amino-Acids   : 46
  CYSS         : 6
 Sequence      : TTCCPSIVARSNFNVCRLPGTPE
                      PDBFINDER2
 DSSP          : CEECSSHHHHHHHHHHHTTTCCH
 Nalign        : 28888888888888888888888
 Nindel        : 00000000000000000010101
 Entropy       : 01002424212511500341433
 Cons-Weight   : 98995354786298299738444
 Cryst-Cont    :     ++++  +   + +++++++
 Access        : 42003687236526724769274
 Quality       :  0.6154
  Present      : 99999999999999999999999
  B-Factors    : 99999999999999999999999
  Bonds        : 98989799998989999999986
  Angles       : 26999969999080898997988
  Torsions     : 08663386233326051124876
  Phi/psi      : ?4544565666666543444564
  Planarity    : 99999999999999999999999
  Backbone     : ??999999999999999909799
  Peptide-Pl   : ??99999999999999999?9?99
  Rotamer      : ??559959954878969599?94
  Chi-1/chi-2: 44455544544757444455455
  Bumps        : 99999999999999999449999
  Packing-1    : 64993976667776540132386
  Packing-2    : 94646644567748554446483
  In/out       : 99999999999999999999999
  H-Bonds      : 99999797999999999999999
  Flips        : 99999999999999999999999
//
```

Figure 4. PDBFINDER and PDBFINDER2 entries of PDB file 1crn (41). The new PDBFINDER2 fields start just below the 'Sequence' field, where the PDBFINDER (40) ends. They provide information about the DSSP secondary structure, the number of aligned UniprotKB sequences, the number of insertions and deletions in these alignments (Nindel), the sequence entropy and conservation weights (all from HSSP). The following fields originate from the PDBREPORTs: Residues involved in crystal contacts, residue accessibilities, and then a large number of structure quality indicators: missing atoms (Present), B-factors, normality of bond lengths, bond angles, torsions, the Ramachandran Plot, side-chain planarity, backbone conformation, peptide-plane orientation, side-chain rotamers,Chi-1/Chi-2 side chain torsion angle distribution, bumps, 3D packing (old and new method) and inside/outside distribution of amino acids. Finally, unsatisfied hydrogen bond donors & acceptors, as well as flipped Asn, Gln and His side-chains are reported. The data are expressed as single digit scores, where most of the time '9' means perfect and '0' terrible, the details can be found at the top of the text file.

not be determined (and which are thus useless for modelling). The PDBFINDER2 provides many more per-residue data aligned with the sequence, which are described in the caption of Figure 4.

In our daily experience, there are two main applications for the PDBFINDERs. The first is complex structure selection queries that cannot be expressed easily in a database language like SQL. For instance, PDBFINDER allows us to quickly select all PDB entries that contain a specific enzyme (by employing the EC number) or all PDB entries that have more than 10 incomplete side chains. The required parsing of the PDBFINDER format takes just a few lines of code, but we also provide a Python module at www.yasara.org/biotools/. The second main application is visualization of the data by mapping it onto the corresponding 3D structure. For this purpose we developed a Python plug-in for the free molecular modelling program 'YASARA View' (42), available from www.yasara.org/viewdl/. Both Python scripts are licensed under the GNU GPL. Figure 5 shows examples of how information from PDBFINDER2 can be visualized.

To study specific properties of proteins structural biologist can study the entire PDB or a representative subset. Such subsets are lists of PDB entries created by filtering the PDB based on criteria of structural uniqueness, structure model quality and experimental parameters. Structural uniqueness is asserted by looking at the pairwise sequence alignment of all entries in the list and setting a cut-off for the maximum allowed sequence identity. From the Sander–Schneider plot, (Figure 2) we see that 25% identity is a safe cut-off. Structure model quality can coarsely be determined by looking at the

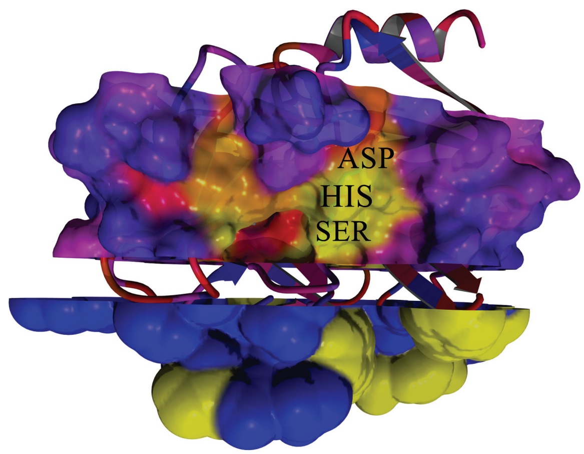

Figure 5. Example of PDBFINDER2 visualization options [2ptn; (43)]. In the bottom slice the solvent accessible surface the surface of residues involved in crystal contacts is coloured yellow. His, Asp, and Ser label the catalytic triad in the slice that shows the molecular surface coloured by HSSP conservation weights (from blue (variable) to yellow (conserved)). The protein backbone is coloured by 3D packing quality in which blue is well packed and red is more poorly packed. Plot made with YASARA (42).

crystallographic (free) R-factor, but a more detailed filter for structure quality uses the results from structure validation software like PROCHECK (44) or WHAT_CHECK. Experimental parameters are usually the type of experiment used to 'solve' the structure (e.g. X-ray crystallography or NMR spectroscopy) and the X-ray resolution. PDBselect by Hobohm and Sander (45–47) provides a good example of methods to select a representative subset of the PDB and so do the PISCES

system (48) and the PDB_REPRDB (49–51). We also have precompiled representative lists of PDB entries in the PDB_SELECT database at http://swift.cmbi.ru.nl/gv/select/ (52). In PDB_SELECT, we have sorted the entries by their quality so that users who take the first N entries from one of the lists will automatically get the best N PDB files where 'best' is defined as a function of resolution, R-factor, and a few WHAT_CHECK quality parameters as described above. Historically, we used a sequence identity cut-off of 30% to balance the requirement of structural uniqueness and getting a large enough data set. With the large increase in size of the PDB, a lower cut-off can be used in future PDB_SELECT sets.

The databases discussed above are kept up-to-date automatically so new entries are continuously added. Sometimes PDB entries are made obsolete rendering their corresponding database entries also obsolete. We developed the WHY_NOT database to keep track of these changes in our other databases. WHY_NOT uses a crawler that runs through a local copy of the PDB and lists which database entries could (in principle) exist and then checks all the databases to see which entries actually do exist, which entries are missing and which entries are obsolete. As the name WHY_NOT implies, the most important function is storing the reasons why certain entries are missing. This serves both the users and maintainers of our databases. For users it is helpful to know that an entry cannot be made and an alternative should be sought, for maintainers it is good to know which entries we should stop trying to make over and over again.

The most trivial reason for a missing database entry is that the PDB entry is so new that corresponding database entries were not created yet. Another simple reason for missing entries is the lack of input data. For instance, PDB_REDO needs the experimental X-ray data; if such data was not deposited, or the structure was solved by other means than X-ray crystallography (such as NMR spectroscopy) a PDB_REDO entry cannot be made. Similarly, a HSSP entry can only be made if a DSSP entry exists. These are obvious reasons for missing entries, but many problems are not straightforward and are annotated in WHY_NOT as 'comments'. For instance DSSP cannot use protein structures that consist only of Cα-atoms, neither can it use PDB entries that contain only nucleic acids or 'other things' such as vancomycin (PDB entry 1sho; (53)). No PDBREPORTs will be made for PDB entries that contain no macromolecules such as PDB entry 1tn1 (54). A PDB_REDO entry cannot be made for X-ray structures in which not all atoms are explicitly listed, but need to be created through matrix operations, which is common practice with viral capsids [e.g. PDB entry 4rhv; (55)]. The most common problems listed in WHY_NOT are given in Table 3.

Most database update procedures add WHY_NOT comments automatically. The update procedure for PDB_REDO is an exception; all WHY_NOT comments are checked by hand. There are two reasons for this: some errors can be traced back to annotation problems in the PDB file or the X-ray data file (e.g. missing R-factors, corrupt TLS group selections, X-ray data stored in the wrong format) and others to limitations in the PDB_REDO software. The PDB_REDO software is topic of ongoing research and is routinely updated to improve dealing with existing PDB problems. Solvable problems in PDB files are always reported to the PDB to ensure that they are fixed at the source rather than by making elaborate workarounds. So far, we have reported some 500 errors in PDB files. Simple administrative problems were fixed swiftly by PDB annotators (typically within two weeks) after which the PDB file was re-released, scientific problems that require information from the depositor and may take longer to be solved.

Table 3. Examples of WHY_NOT comments

Database	Comment	Occurrences in the PDB
DSSP	Nucleic acids only	2.1k
HSSP	No alignable sequence	97
PDBREPORT	Too many C-α only residues	211
PDB_REDO	No R-factor reported	66

INTEGRATED UPDATING MECHANISM

All but one of PDB derived databases are updated with every new PDB release (PDB_SELECT is updated annually or upon request). When a new entry is added to the PDB or an existing entry is altered, its corresponding database entries are also (re)created. Our databases are interdependent via 'hard' dependencies (e.g. no HSSP entry can be made without a DSSP entry) and 'soft' dependencies (PDB_REDO uses PDBREPORT if an entry is available). The dependencies between databases are depicted in Figure 6. The process of building our databases resembles building software from source code where one creates object files out of source files, which are then linked into executables. Because of this similarity we have chosen the ubiquitous make to do the actual work and the rules are written in Makefiles and the result is a very flexible and robust system. Once a week, the make process is started by a 'cron' job and then it starts fetching the latest updates for PDB. After updating PDB, the depending databanks are built, guided by the Makefiles and the dependencies embodied therein. We have tweaked the Makefiles to allow for an exception for replacing existing HSSP files: HSSP uses the UniprotKB database, but because UniprotKB and PDB entries do not map 1-to-1, 'all' HSSP entries should be updated with every new release of the Uniprot knowledge base. This makes the maintenance of HSSP files a quadratic problem because each PDB entry is aligned against all UniProtKB entries, and both databases grow continuously. We do not have the CPU power available to update all HSSP files at every UniprotKB release; instead we update as many HSSP files older than 6 months as we can (typically a few thousand) with the remaining CPU time of our 1 week update cycle.

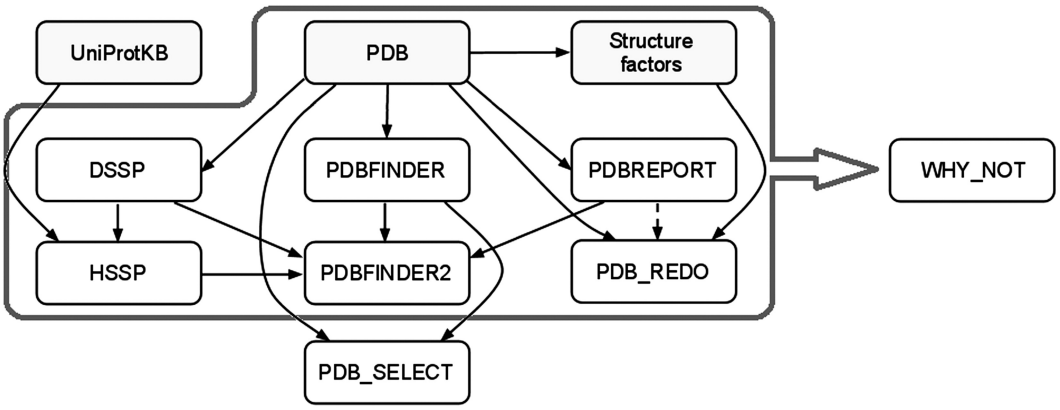

Figure 6. Database dependency schema. Dependencies between our databases (white background boxes) and three data sources (gray background boxes). A solid arrow means that an entry can only be made if an entry in the box where the arrow comes from exists. A dotted arrow means that data is used when available. The databases inside the gray line are indexed in WHY_NOT.

Table 4. Rsync access to the databases

Databank	Access
DSSP	rsync -avz rsync://rsync.cmbi.ru.nl/dssp/ dssp/
HSSP	rsync -avz rsync://rsync.cmbi.ru.nl/hssp/ hssp/
PDBFINDER	rsync -avz rsync://rsync.cmbi.ru.nl/pdbfinder/ pdbfinder/
PDBFINDER2	rsync -avz rsync://rsync.cmbi.ru.nl/pdbfinder2/ pdbfinder2/
PDB_REDO	rsync -avz rsync://rsync.cmbi.ru.nl/pdb_redo/ pdb_redo/

ACCESS

Multiple forms of access to the systems exist. The MRS system (mrs.cmbi.ru.nl) is a generic, freely available database query system that has been described elsewhere (39). MRS provides access to about 60 international databases that we use often enough to warrant in-house shadowing. MRS can also be used to query all databases mentioned in this article, except WHY_NOT. MRS also handles Web service requests, either using SOAP or the REST protocol. Five of the systems can be shadowed in-house using the rsync protocols listed in Table 4.

WHY_NOT is accessible via the WHY_NOT query system. DSSP can additionally be accessed through the WHAT IF Web servers (swift.cmbi.ru.nl) or through the WIWS Web services (WSDL address: http://wiws.cmbi.ru.nl/wsdl); these two systems also allow the user to upload his/her own PDB file for secondary structure determination.

PDB_REDO and PDBREPORT are also directly linked at every entry page of the EBI interface of the PDB.

FUTURE WORK

We continue to work on our databases in order to improve the quality and usability. An improvement of quality comes mostly from adding new options to the WHAT_CHECK software and the PDB_REDO pipeline. Both are subject of ongoing research and new features are added frequently. The PDBREPORT database will be completely rebuilt when a new WHAT_CHECK is released by the end of 2010. We are also working on improving our software to reduce the number of missing entries or, if all else fails, have clear explanations why certain entries cannot be made. Our WHY_NOT database will be an important resource to achieve this.

In terms of usability, we are working on making our databases easier to access. For instance, PDBREPORT can be indexed by our MRS database searching software. PDB_REDO structures will be accessible directly from molecular viewers such as YASARA. We are also working on new dissemination tools to guide the user in using our databases. We focus strongly on visualization: the WHAT_CHECK user course currently under development has numerous visual examples of the warnings and errors that can be found in PDBREPORT. The latest version of the PDB_REDO pipeline creates YASARA scenes that show exactly which atoms moved the most when a PDB entry was optimized.

ACKNOWLEDGEMENTS

The authors are especially grateful to those users of these databases who cite the related articles and who report problems. Each of the databases holds between fifty thousand and sixty seven thousand entries, and it is impossible to always have everything correct and up-to-date. User support is highly valued. The EU contributed in the 90′s to the initial design of several of the systems mentioned. More recent financial support came also from EMBRACE, BioSapiens, Elixir and NBIC.

FUNDING

Funding for open access charge: National Institutes of Health (Grant R01 GM62612); the Stichting Nationale Computerfaciliteiten (National Computing Facilities Foundation, NCF) for the use of supercomputer facilities; Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organization for Scientific Research, NWO).

Conflict of interest statement. None declared.

REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
3. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
4. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
5. The UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res*, **36**, D190–D195.
6. Hekkelman,M.L., Te Beek,T.A.H., Pettifer,S.R., Thorne,D., Attwood,T.K. and Vriend,G. (2010) WIWS: a protein structure bioinformatics Web service collection. *Nucleic Acids Res.*, **38(Suppl)**, W719–W723.
7. Joosten,R.P. and Vriend,G. (2007) PDB improvement starts with data deposition. *Science*, **317**, 195–196.
8. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
9. Yang,J., Peng,Z. and Chen,X. (2010) Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. *BMC Bioinformatics*, **11(Suppl. 1)**, S9.
10. Madera,M., Calmus,R., Thiltgen,G., Karplus,K. and Gough,J. (2010) Improving protein secondary structure prediction using a simple k-mer model. *Bioinformatics*, **26**, 596–602.
11. Babaei,S., Geranmayeh,A. and Seyyedsalehi,S.A. (2010) Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. *Comput. Methods Programs Biomed.*, **100**, 237–247.
12. Chen,K., Stach,W., Homaeian,L. and Kurgan,L. (2010) iFC(2): an integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content. *Amino Acids*, [doi;10.1007/s00726-010-0721-1, Epub ahead of print 21 Aug 2010].
13. Pirovano,W. and Heringa,J. (2010) Protein secondary structure prediction. *Methods Mol. Biol.*, **609**, 327–348.
14. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
15. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
16. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
17. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
18. Sander,C. and Schneider,R. (1994) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **22**, 3597–3599.
19. Schneider,R. and Sander,C. (1996) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **24**, 201–205.
20. Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
21. Dodge,C., Schneider,R. and Sander,C. (1998) The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.*, **26**, 313–315.
22. Chang,G., Roth,C.B., Reyes,C.L., Pornillos,O., Chen,Y. and Chen,A.P. (2006) Retraction. *Science*, **314**, 1875.
23. Vriend,G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph*, **8**, 52–56, 29.
24. Matthews,B.W. (1968) Solvent content of protein crystals. *J. Mol. Biol.*, **33**, 491–497.
25. Cremer,D. and Pople,J.A. (1975) General definition of ring puckering coordinates. *J. Am. Chem. Soc.*, **97**, 1354–1358.
26. Engh,R.A. and Huber,R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr. Sect. A*, **47**, 392–400.
27. Vriend,G. and Sander,C. (1993) Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallogr.*, **26**, 47–60.
28. Hooft,R.W.W., Sander,C. and Vriend,G. (1994) Reconstruction of symmetry-related molecules from protein data bank (PDB) files. *J. Appl. Crystallogr.*, **27**, 1006–1009.
29. Parkinson,G., Vojtechovsky,J., Clowney,L., Brünger,A.T. and Berman,H.M. (1996) New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallogr. D Biol. Crystallogr.*, **52**, 57–64.
30. Hooft,R.W., Sander,C. and Vriend,G. (1996) Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins*, **26**, 363–376.
31. Hooft,R.W., Sander,C. and Vriend,G. (1997) Objectively judging the quality of a protein structure from a Ramachandran plot. *Comput. Appl. Biosci.*, **13**, 425–430.
32. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.*, **53**, 240–255.
33. Winn,M.D., Isupov,M.N. and Murshudov,G.N. (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D Biol. Crystallogr.*, **57**, 122–133.
34. Joosten,R.P., Salzemann,J., Bloch,V., Stockinger,H., Berglund,A., Blanchet,C., Bongcam-Rudloff,E., Combet,C., Da Costa,A.L., Deleage,G. *et al.* (2009) PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J. App. Crystallogr.*, **42**, 376–384.
35. Roe,S.M., Barlow,T., Brown,T., Oram,M., Keeley,A., Tsaneva,I.R. and Pearl,L.H. (1998) Crystal structure of an octameric RuvA-Holliday junction complex. *Mol. Cell*, **2**, 361–372.
36. Schomaker,V. and Trueblood,K.N. (1968) On the rigid-body motion of molecules in crystals. *Acta Crystallogr. Sect.B*, **24**, 63–76.
37. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Meth. Enzymol.*, **266**, 114–128.
38. Valentin,F., Squizzato,S., Goujon,M., McWilliam,H., Paern,J. and Lopez,R. (2010) Fast and efficient searching of biological data resources–using EB-eye. *Brief. Bioinformatics*, **11**, 375–384.
39. Hekkelman,M.L. and Vriend,G. (2005) MRS: a fast and compact retrieval system for biological data. *Nucleic Acids Res.*, **33**, W766–W769.
40. Hooft,R.W., Sander,C., Scharf,M. and Vriend,G. (1996) The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput. Appl. Biosci.*, **12**, 525–529.
41. Teeter,M.M. (1984) Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. *Proc. Natl Acad. Sci. USA*, **81**, 6014–6018.
42. Krieger,E., Koraimann,G. and Vriend,G. (2002) Increasing the precision of comparative models with YASARA NOVA–a self-parameterizing force field. *Proteins*, **47**, 393–402.
43. Walter,J., Steigemann,W., Singh,T.P., Bartunik,H., Bode,W. and Huber,R. (1982) On the disordered activation domain in trypsinogen: chemical labelling and low-temperature crystallography. *Acta Cryst. Sect. B*, **38**, 1462–1472.
44. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. App. Crystallogr.*, **26**, 283–291.
45. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.
46. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.
47. Griep,S. and Hobohm,U. (2010) PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res.*, **38**, D318–D319.
48. Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
49. Noguchi,T., Onizuka,K., Akiyama,Y. and Saito,M. (1997) PDB-REPRDB: a database of representative protein chains in

PDB (Protein Data Bank). *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 214–217.
50. Noguchi,T., Matsuda,H. and Akiyama,Y. (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res.*, **29**, 219–220.
51. Noguchi,T. and Akiyama,Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res.*, **31**, 492–493.
52. Hooft,R.W.W., Sander,C. and Vriend,G. (1996) Verification of protein structures: side-chain planarity. *J. App. Crystallogr.*, **29**, 714–716.
53. Schäfer,M., Schneider,T.R. and Sheldrick,G.M. (1996) Crystal structure of vancomycin. *Structure*, **4**, 1509–1515.
54. Brown,R.S., Dewan,J.C. and Klug,A. (1985) Crystallographic and biochemical investigation of the lead(II)-catalyzed hydrolysis of yeast phenylalanine tRNA. *Biochemistry*, **24**, 4785–4801.
55. Arnold,E. and Rossmann,M.G. (1988) The use of molecular-replacement phases for the refinement of the human rhinovirus 14 structure. *Acta Crystallogr., A, Found. Crystallogr.*, **44(Pt 3)**, 270–282.